TEMPORAL ASYNCHRONICITY MODELING BY PRODUCT HMMS FOR AUDIO-VISUAL SPEECH RECOGNITION

Satoshi Nakamura

ATR Spoken Language Translation Research Laboratories
satoshi.nakamura@atr.co.jp

ABSTRACT

There have been higher demands recently for Automatic Speech Recognition (ASR) systems able to operate robustly in acoustically noisy environments. This paper proposes a method to effectively integrate audio and visual information in audio-visual (bi-modal) ASR systems. Such integration inevitably necessitates modeling of the synchronization and asynchronization of the audio and visual information. To address the time lag and correlation problems in individual features between speech and lip movements, we introduce a type of integrated HMM modeling of audio-visual information based on a family of a product HMM. The proposed model can represent state synchronicity not only within a phoneme but also between phonemes. Furthermore, we also propose a rapid stream weight optimization based on GPD algorithm for noisy bi-modal speech recognition. Evaluation experiments show that the proposed method improves the recognition accuracy for noisy speech. In SNR=0dB our proposed method attained 16% higher performance compared to a product HMMs without the synchronicity re-estimation.

1. INTRODUCTION

The performance of ASR systems has been drastically improved recently. However, it is well known that the performance can be seriously degraded in acoustically noisy environments. Audio-visual ASR [1, 2, 4] systems offer the possibility of improving the conventional speech recognition performance by incorporating visual information, since the speech recognition performance is always degraded in acoustically noisy environments whereas visual information is not.

Audio and visual phonetic features have different durations. In other words, there is loose synchronicity between them, for instance, a speaker opens the mouth before making an utterance, and closes it after making the utterance. Furthermore, the time lag between the movement of the mouth and the voice might be dependent on the speaker or context.

As audio-visual integration methods for ASR systems, early integration and late integration are well known [1, 2]. In the early integration scheme, a conventional HMM is trained using audio-visual data. This method, however, cannot sufficiently represent the loose synchronization between the audio and visual information. Furthermore, the visual features of the conventional HMM may end up relatively poorly trained because of mis-alignments during the model estimation caused by the segmentation of the audio features. In the late integration scheme, the audio data and visual data are processed separately to build two independent HMMs [1, 4]. This scheme assumes complete asynchronization between the audio and visual features. In addition, it can make the best use of the audio and visual data because there is a smaller bi-modal database than the typical database for audio only. However, the audio and visual features are regarded as independent. In this paper, in order to model the synchronization between audio and visual features, we propose pseudo-biphone product HMMs which realizes state synchronous audio-visual integration. The proposed model can represent synchronicity not only within a phoneme but also beyond phoneme boundaries. Furthermore, we propose a new method based on GPD algorithm to optimize stream weights of the proposed pseudo-biphone product HMMs.

2. AUDIO-VISUAL INTEGRATION BASED ON PRODUCT HMM

Figure 1 shows the outline of the acoustic model training for ASR systems in this paper. Figure 2 shows the proposed HMM topology. First, in order to create the audio and visual phoneme HMMs independently, audio features and visual features are extracted from audio data and visual data, respectively. In general, the frame rate of audio features is higher than that of visual features. Accordingly, the extracted visual features are incorporated such that the audio and visual features have the same frame rate. Second, the audio and visual features are modeled individually into two HMMs by the EM algorithm. Finally, an audio-visual phoneme HMM is composed as the product of these two HMMs based on HMM composition. The output probability at state ij of the audio-visual HMM is,

$$b_{ij}(O_t) = b_i^A(O_t^A)^{\alpha_A} \times b_j^V(O_t^V)^{\alpha_V} \quad (1)$$

which is defined as the product of the output probabilities of the audio and visual streams. Here, $b_i^A(O_t^A)^{\alpha_A}$ is the output probability

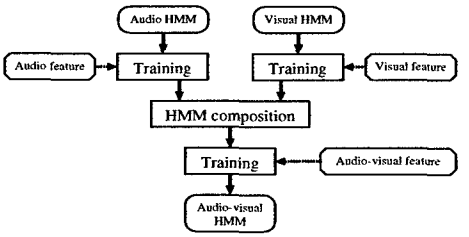

Fig. 1. Procedure Overview

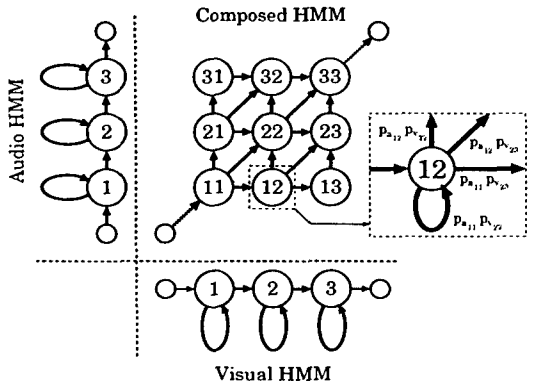

Fig. 2. Product HMM

of the audio feature vector at time instance t in state i, $b_j^V(O_t^V)^{\alpha_V}$ is the output probability of the visual feature vector at time instance t in state j, and α_A and α_V are the audio stream weight and visual stream weight, respectively. In a similar manner, the transition probability from state ij to state kl in the audio-visual HMM is defined as follows.

$$p_{ij.kl} = p_{a_{i.k}} \times p_{v_{j.l}} \quad (2)$$

where $p_{a_{i.k}}$ is the transition probability from state i to state k in the audio HMM, and $p_{v_{j.l}}$ is the transition probability from state j to state l in the visual HMM. This composition is performed for all phonemes. In the method proposed by [4], a similar composition is used for the audio and visual HMMs. However, because the audio and visual HMMs are trained individually, the dependencies between the audio and visual features are ignored. This results in the following two problems.

1. The product HMMs can not represent the loose synchronicity within phonemes as it is.
2. The product HMMs force a strict synchronization on every phoneme boundary.

This paper proposes a new approach to solve the two problems. The approach proposes re-estimation of the product HMMs parameters by using a small amount of audio-visual synchronous adaptation data, and pseudo-biphone product HMMs which represent loose state synchronicity beyond the phoneme boundary.

2.1. State Synchronous Modeling within a Phoneme

The first problem is from the inability of the conventional product HMMs to represent loose state synchronicity within a phoneme. This problem is caused by the fact that the transition probabilities and output probabilities are obtained by the multiplication of probabilities from independent states of audio and visual HMMs. We propose new product HMMs whose parameters are re-estimated using audio-visual synchronous adaptation data [3]. The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMM. The re-estimation procedure is carried out using a small amount of audio-visual synchronous data. After the composition of two HMMs, the product HMMs can be re-estimated based on the Baum-Welch algorithm for multi-stream HMMs.

Figure 3 shows results comparing audio HMMs, visual HMMs, early integration, late integration, and product HMMs with and without re-estimation [3]. The experimental conditions are the same as those in a later section except that the audio HMMs are trained using clean speech data. The figure shows that the product HMMs with re-estimation achieve the best performance, while the product HMMs without re-estimation are worse than those of the early and late integration schemes.

2.2. State Synchronous Modeling Beyond The Phoneme Boundary

The second problem is that the conventional product HMMs force a strict synchronization on every phoneme boundary. This is because the speech organs normally move earlier than the speech to be produced. Sometimes, the speech organs are already articulated in the previous audio phoneme utterance. Accordingly, we have to consider state synchronous modeling beyond the phoneme boundary. We have carried out preliminary experiments using audio-visual word HMMs and confirmed that synchronicity is not always kept on a phoneme boundary looking at the optimal paths[5].

We propose new product HMMs that include extra asynchronous states on phoneme boundaries as indicated in Fig. 4. The core states of the phoneme HMMs are the same as those of context independent phoneme product HMMs. In addition, the new product HMMs have two extra HMM states aiming to work similarly to the word HMMs. The first extra state is composed of the initial audio state and final visual state of the preceding phoneme HMM. The second extra state is composed of the initial visual state and final audio state of the preceding phoneme HMM. Since these extra states are dependent on the preceding phoneme, they can only be re-estimated in a manner similar to the biphone HMMs. Therefore, we call these HMM pseudo-biphone product HMMs. The proposed HMMs can tolerate one state asynchronicity beyond a phoneme boundary.

3. STREAM WEIGHT OPTIMIZATION

As methods for estimating stream weights, maximum likelihood [6] based methods or GPD (Generalized Probabilistic Descent)[7] based methods have been proposed. However, the former methods have a serious estimation drawback because the scales of two probability are normally very different and so the weights can not be estimated optimally. The latter methods have substantial possibility for optimizing the weights. However, a serious problem is that these methods require a lot of adaptation data is necessary

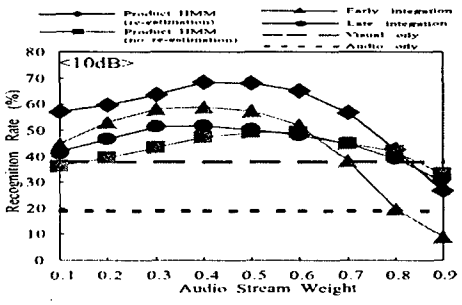

Fig. 3. Results of Product HMMs

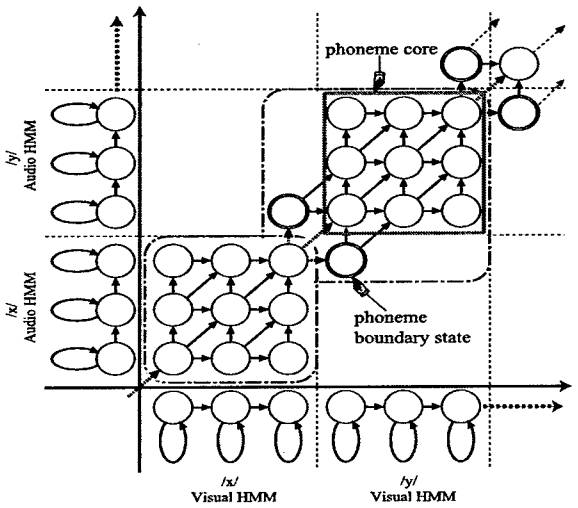

Fig. 4. Pseudo-biphone product HMMs

for the weight estimation. In this paper, we propose a GPD-based simplified adaptive estimation of stream weights using GMMs for new noisy acoustic conditions.

The approach by the GPD training defines a misclassification measure, which provides distance information concerning the correct class and all other competing classes. The misclassification measure is formulated as a smoothed loss function. This loss function is minimized by the GPD algorithm. Here, let $L_c^{(x)}(\Lambda)$ be the log-likelihood score in recognizing input data x for adaptation using the correct word model, where $\Lambda = \{\lambda_A, \lambda_V\}$.

In a similar way, let $L_n^{(x)}(\Lambda)$ be the score in recognizing data x using the n-th best candidate among the mistaken word models.

The misclassification measure is defined as,

$$d^{(x)} = -L_c^{(x)}(\Lambda) + \log[\frac{1}{N}\sum_{n=1}^{N}\exp\{\eta L_n^{(x)}(\Lambda)\}]^{\frac{1}{\eta}} \quad (3)$$

where η is a positive number, and N is the total number of candidates. The smoothed loss function for each data is defined as,

$$l^{(x)} = [1+\exp\{-\alpha d^{(x)}(\Lambda)\}]^{-1} \quad (4)$$

where α is a positive number. In order to stabilize the gradient, the loss function for the entire data is defined as,

$$l(\Lambda) = \sum_{x=1}^{x} l^{(x)}(\Lambda) \quad (5)$$

where X is the total amount of data. The minimization of the loss function expressed by equation (5) is directly linked to the minimization of the error. The GPD algorithm adjusts the stream weights recursively according to,

$$\Lambda_{k+1} = \Lambda_k - \varepsilon_k E_k \nabla l(\lambda), k = 1, .., \quad (6)$$

where $\varepsilon_k > 0, \sum_{k=1}^{\infty}\varepsilon_k = \infty, \sum_{k=1}^{\infty}\varepsilon_k^2 < \infty$, and E is a unit matrix.

In this paper, we propose to use GMMs instead of HMMs to find optimal stream weights not for the recognition. GPD training on GMMs is quite simple and requires smaller amount of training data. We use 18 mixture Gaussians for GMMs and train them using all of the training data.

4. EVALUATION EXPERIMENTS

The audio signal is sampled at 12 kHz (down-sampled) and analyzed with a frame length of 32 msec every 8 msec. The audio features are 16-dimensional MFCC and 16-dimensional delta MFCC. On the other hand, the visual image signal is sampled at 30 Hz with 256 gray scale levels from RGB. Then, the image level and location are normalized by a histogram and template matching. Next, the normalized images are analyzed by two-dimensional FFT to extract 6x6 log power 2-D spectra for audio-visual ASR. Finally, 35-dimensional 2D log power spectra and their delta features are extracted. For each modality, the basic coefficients and the delta coefficients are collectively merged into one stream. Since the frame rate of the video images is 1/30, we insert the same images so as to synchronize the face image frame rate to the audio speech frame rate. For the HMMs, we use a two-mixture Gaussian distribution and assign three states for the audio stream and two states for the visual stream in the late integration HMMs and the baseline product HMMs. In this research, we perform word recognition evaluations using a bi-modal database [1]. We use 4740 words for HMM training and two sets of 200 words for testing. These 200 words are different from the words used in the training. We perform experiments using 15, 25, and 50 words. The context of the data for the adaptation differs from that of the test data. In order to examine in more detail the estimation accuracy in the case of less adaptation data, we carry out recognition experiments using three sets of data, each as different as possible from the context. The size of the vocabulary in the dictionary is 500 words during the recognition of the adaptation data. The GPD algorithm convergence pattern is known to greatly depend on the choice of parameters. Accordingly, we set $N = 1$ in (3), $N = 0.1$ in (4), $N = 100/k$, and the maximum the iteration count $= 8$.

We compared the processed product HMMs without re-estimation (Product-HMM(W/O Re-est.)), the proposed product HMMs with re-estimation (Product-HMM(W Re-est.)), the proposed pseudo-biphone product HMMs without re-estimation (Pseudo-Biphon(W/O Re-est.)), the proposed pseudo-biphone product HMMs with re-estimation (Pseudo-Biphon(W Re-est.)), and GMM for GPD-based stream weight optimization for acoustic SNR=15, 0, and -5dB. White noise was used to reduce the acoustic SNR in this experiment. The audio HMMs were trained using the SNR=15dB data. The results indicate that the re-estimation of the product HMMs is quite effective to improve the performance. The re-estimation is able to introduce the loose state synchronicity of the states of two modalities into the product HMMs. The state synchronous modeling beyond the phoneme boundary by a pseudo-biphone product HMM also results in significant improvements to the product HMMs. It is also confirmed that the re-estimation further improves performance of pseudo-biphone product HMMs. The figures show optimal stream weights for the maximum performance vary according to each method and acoustic SNR. The solid arrows show the results by simplified GPD-based stream weight estimation using 25 adaptation words. The proposed GPD-based simplified stream weight optimization algorithm successfully estimated stream weight with almost the best performance. In the SNR=-5dB environment, the estimated weight is not the optimal one. Figure 8 shows standard deviation of the word accuracy over various SNRs, a number of adaptation words, and a number of candidates in GPD training. It is confirmed the standard deviation in SNR=-5dB is bigger than the others and smaller number of adaptation words gives bigger standard deviations. In SNR=0dB our

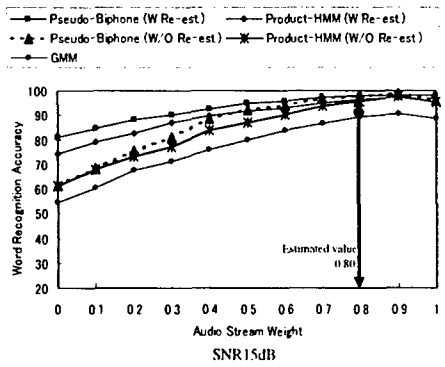

Fig. 5. Word Accuracy (SNR=15dB)

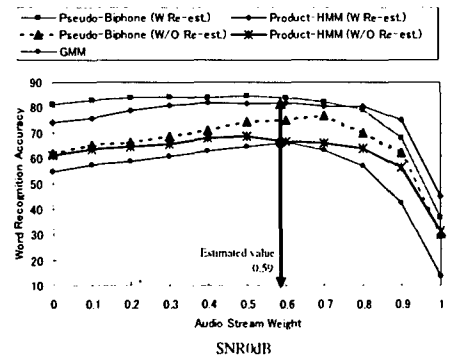

Fig. 6. Word Accuracy (SNR=0dB)

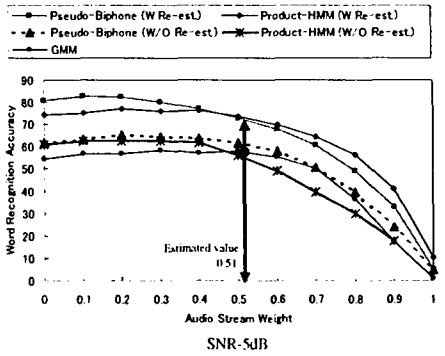

Fig. 7. Word Accuracy (SNR=-5dB)

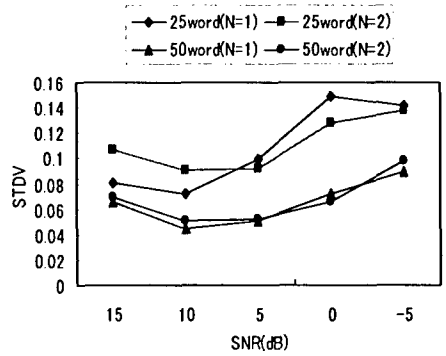

Fig. 8. Standard Deviation of Word Accuracy

proposed method attained 16% higher performance compared to a product HMMs without the synchronicity re-estimation.

5. CONCLUSION

This paper proposes a new HMM structure to effectively integrate audio and visual information in audio-visual (bi-modal) systems. Our state synchronous modeling of audio-visual information is based on the product HMM. The proposed model can represent synchronicity not only within a phoneme but also between phonemes. Evaluation experiments show that the re-estimation of the model parameters using audio-visual synchronous data further improves the product HMMs. In addition, pseudo-biphone HMMs that introduce two extra asynchronous states are shown to improve the bimodal speech recognition accuracy. Furthermore, we also proposed a rapid stream weight optimization based on GPD algorithm for noisy bi-modal speech recognition.

6. ACKNOWLEDGEMENTS

The authors thank intern students, K.Kumatani and S.Tamura, and their supervisors, Prof. S. Furui of the Tokyo Institute of Technology and Prof. K. Shikano of the Nara Institute of Science and Technology for giving us the opportunity to conduct this study collaboratively. This research was supported in part by the Telecommunications Advancement Organization of Japan.

7. REFERENCES

[1] S.Nakamura,et al., "Improved bimodal speech recognition using tied-mixture HMMs and 5000 word Audio-Visual Synchronous database", Proc. Eurospeech97

[2] S.Nakamura, et al., "Stream weight optimization of speech and lip image sequence for Audio-Visual speech recognition", Proc. ICSLP2000

[3] Kenichi Kumatani, Satoshi Nakamura and Kiyohiro Shikano, "An Adaptive Integration Method Based on Product HMM for Bi-Modal Speech Recognition", HSC2001 (International Workshop on Hands-Free Speech Communication) pp. 195-198

[4] M.J. Tomlinson, et al.,"Integrating audio and visual information to provide highly robust speech recognition", Proc. ICASSP-96

[5] S.Nakamura, K.Kumatani, S.Tamura, "State Synchronous Modeling of Audio-Visual Information for Bi-modal Speech Recognition", Proc. IEEE ASRU Dec. 2002

[6] J.Hernando," Maximum Likelihood Weighting of Dynamic Speech Features for CDHMM Speech Recognition", Proc. ICASSP'97,(1997) 1267–1270

[7] G.Potamianos, H.P.Graf, "Discriminative Training of HMM Stream Exponents for Audio-visual Speech Recognition", Proc. ICASSP'98,(1998) 3733–3736